图书情报工作
第65卷 第16期 2021年8月

基于知识图谱的红色历史人物知识问答服务框架研究

■ 张云中 郭冬 王亚鸽 孙平
上海大学图书情报档案系 上海 200444

摘 要：[目的/意义]知识图谱已成为公共数字文化资源知识组织的新形态。利用知识图谱技术赋能红色历史人物知识问答服务，提升用户交互体验，对红色历史资源的开发利用具有重要意义。[方法/过程]在梳理历史人物数字资源组织及知识问答系统相关研究的基础之上，构建了红色历史人物知识图谱 Schema 与 KBQA 架构，从数据获取、知识抽取、知识融合、图谱生成和知识问答五个环节搭建了红色历史人物问答模型，并以老上大历史人物数字资源进行实证研究。[结果/结论]本文设计的知识问答服务架构在红色历史人物数字资源的图谱半自动构建、知识推理与智能交互方面具有优越性，提升了用户知识服务体验。

关键词：红色历史人物 知识图谱 问答系统 知识服务
分类号：G203
DOI：10.13266/j.issn.0252-3116.2021.16.012

1 引言

红色文化具有鲜明的中国风格和典型的中国特色，它在长期革命、建设和发展的历史进程中，产生并留存了大量不同形式、样态的红色文化资源。其中，红色历史人物作为红色文化的缔造者和传播者，是红色文化资源内容呈现和展演的主要载体。近年来，习近平总书记反复强调“要把红色资源利用好、把红色传统发扬好、把红色基因传承好”，推进红色文化资源尤其是红色历史人物资源的数字化组织、管理和开发利用，对推广红色历史人物，弘扬红色文化精神有着重要的理论价值和实践意义。中共中央办公厅、国务院办公厅曾印发《关于实施革命文物保护利用工程（2018—2022年）的意见》[1]，文件中强调要适度运用现代科技手段，增强革命文物陈列展览的互动性、体验性，真正让红色文化“活起来”，充分挖掘红色文化资源内涵，提高红色文化数字资源库的利用率。

目前，红色历史人物数字资源的组织、管理和利用模式已产生了“从数据孤立到数据关联”的变革。知识图谱作为当下新兴且应用广泛的一种展现形式，能够提供语义化、可视化、智慧化的数字资源知识组织范式，将其应用至红色历史人物数字资源上，不仅有助于重现并挖掘红色历史人物资源中的知识关联，为展现红色历史人物中以人物经历和关系为脉络的各类信息内容提供了可能，同时能够极大程度丰富红色历史人物资源的知识发现等服务方式，进一步深化红色历史人物数字资源的开发利用。

因此，本研究拟以知识图谱技术，结合知识问答服务框架，深入探讨红色历史人物数字资源组织、管理和开发利用的新方式，以期实现红色历史人物关联数据的发布，从而完善红色历史人物相关的数据基础设施建设，并为党史馆、博物馆等红色旅游遗址及文创产品的建设、研发奠定基础。

1 研究现状

本研究对于红色历史人物数字资源的组织、管理和开发利用实则建构于两个核心问题之上：历史人物数字资源组织及知识服务的研究现状如何？以知识图谱为基础进行问答服务的方法论研究现状如何？下文述评围绕上述两个关键点展开。

1.1 历史人物数字资源组织及知识服务研究现状

近年来，历史人物数字资源组织及知识服务领域的研究主要包括以下三个方面：一是历史人物数据库的建设，典型数据库有古代人物关系数据库“中国历代

作者简介：张云中（ORCID：0000-0002-7323-2561），副教授，博士，硕士生导师，E-mail：zhang-yun-zhong@126.com；郭冬（ORCID：0000-0003-2462-0071），硕士研究生；王亚鸽（ORCID：0000-0003-2107-4263），硕士研究生；孙平（ORCID：0000-0003-1681-5397），硕士研究生。
收稿日期：2021-03-29 修回日期：2021-06-07 本文起止页码：108-117 本文责任编辑：杜杏叶

人物传记资料库(CBDB)”[2]、文化部全国文化信息资源共享工程“湖南近代人物资源库”[3]、河北红色历史文化资源“李大钊专题数据库”[4]等,为历史人物数据库的概念分类、层级搭建、内容选取等研究提供了经验;二是以历史人物数据库为数据基础的数据拓展与演变,其代表有基于RDF形式化描述的学术名人知识模型[5]、中国历代人文传记资料库关联数据平台(CB-DBLD)[6]、CBDB历史人物关系网络[7]等,皆为历史人物资源数据的形式化表达和语义关联奠定了基础;三是以人物关系可视化展示、知识问答为代表的知识服务,其成果有宋代学术师承关系可视化展示[8]、历史人物实体关系可视化系统[9]和中国历史人物知识智能问答系统[10]等,为历史人物的知识服务研究拓宽了思路。以上研究均以“历史人物数字资源”为研究对象,推进了历史人物数字资源知识组织及知识服务的研究进程,为本研究奠定了数据及技术基础。

1.2 基于知识图谱的知识问答服务方法论研究现状

基于知识图谱的问答服务是目前知识问答领域的热点,根据众多领域知识内容的不同,可将问答服务实现方法粗分为如下四类:第一类,基于模板匹配的问答方法,其关键在于预设SPARQL模板[11],进而依据问题类型选取模板生成答案,代表服务如疾病问答系统[12]、投资问答系统[13]等;第二类,基于语义解析的问答方法,其模式为通过解析自然语言问句来返回相应结果,如中国历史人物知识智能问答系统、馆藏文物资源知识关联与智能问答系统[14]等;第三类,基于深度学习的问答方法,其思路为通过神经网络等技术优化问答模型,代表研究如LSTM神经网络构建的问答模型[15]、基于BERT和BiLSTM-CRF的古诗知识图谱智能问答系统[16];第四类,基于知识推理的智能问答方法,其思路是通过路径推理计算得到知识图谱中的隐含知识,如基于多模态信息循环推理的知识问答系统[17]、采用MHRP的知识推理框架[18]等。上述研究从多方面阐述了基于知识图谱的知识问答服务构建方法,均对本研究有着借鉴意义。

统而言之,目前就历史人物数字资源组织、管理与开发利用的相关研究虽已取得了较为丰富的成果,但聚焦到基于知识图谱的问答系统构建研究上仍然存在不足之处:一是知识图谱构建的数据源多以结构化数据为主,鲜有从半结构化数据出发构建图谱的尝试;二是针对历史人物知识图谱的Schema设计在与用户需求相匹配上存在不足;三是基于知识图谱的知识问答服务架构有待优化,特别是意图识别和知识推理方法有待进一步完善。本研究将结合红色历史人物知识问答服务的两大关键,着力解决上述不足问题。

2 基于知识图谱的红色历史人物知识问答服务框架设计

2.1 红色历史人物知识问答服务的两个关键问题

2.1.1 知识库设计:红色历史人物知识图谱Schema

本文主要采用体验式设计法设计红色历史人物知识库,旨在将用户知识问答服务核心需求与红色历史人物知识图谱设计真正匹配起来,以增强用户的交互体验。研究招募红色历史人物兴趣爱好者二十名,从红色文献相关的百度百科、数据库、微信公众号等数据源选取具有代表性的红色历史人物资料50余篇,通过兴趣爱好者阅读材料、析取兴趣问题、问题分类聚焦等过程,将其对所关注历史人物的问答需求聚焦为基本信息、革命履历、作品著述、社会关系、档案资源等5个方面。根据以上问题需求,然后结合已有历史人物资料库的信息内容,按照自顶向下的方式,从中析取主要实体12类、关系23类、关键属性4类,并设计出红色历史人物知识图谱的Schema,见图1。

红色历史人物知识图谱的Schema要素主要涵盖实体、属性、关系三个类别,其中实体主要揭示与红色历史人物相关的客观个体,如人物姓名、代表作名称、事件名称和地点名称等;属性是对实体内涵的结构化描述,如事件的背景和影响等;关系主要揭示实体与实体之间所蕴含的某种联系,如“人物”和“人物”之间可以是“师生/夫妻/同学/亲友”等关系。通过实体-属性-描述、实体-关系-实体的三元组框架,红色历史人物知识节点之间的网络关联得以建立,并可用RDF数据格式形式化表达。需要强调的是,红色历史人物知识图谱的Schema设计原则是简洁有效,对于无法通过三元组数据直接获取的信息,可通过知识推理实现。以图1所示虚线关系为例,通过对知识图谱中人物革命履历进行数据挖掘可以分析出人物与事件或人物与地名的隐含联系,通过对图谱中不同人物之间的关系进行多步计算则可以推理得到某些人物之间隐含的关系,这些都是对知识图谱已有数据进行挖掘、分析和推理之后,从已知事实出发找出其中所蕴含新“知识”的过程。

另外,红色历史人物知识图谱的Schema设计还需遵循开放关联数据原则,并尽可能减少数据规约,从而使其易于与场外新数据进行融合,为数据增广和长路

径推理提供概念基础。目前,红色历史人物知识图谱 Schema 主要涵盖红色历史人物的基本信息、革命履历、作品著述、社会关系、档案资源等多方位信息,其设计可随着需求变化和时间推移不断按需动态扩充,只需将新加入的实体、关系、属性按照关联数据标准链接到现有 Schema 即可。红色历史人物知识图谱 Schema 构建为红色历史人物的知识组织提供了底层的数据基础设施支撑,便于据此构建红色历史人物知识图谱,进而为 KBQA 框架设计中的问答 Agent 提供应用支撑。

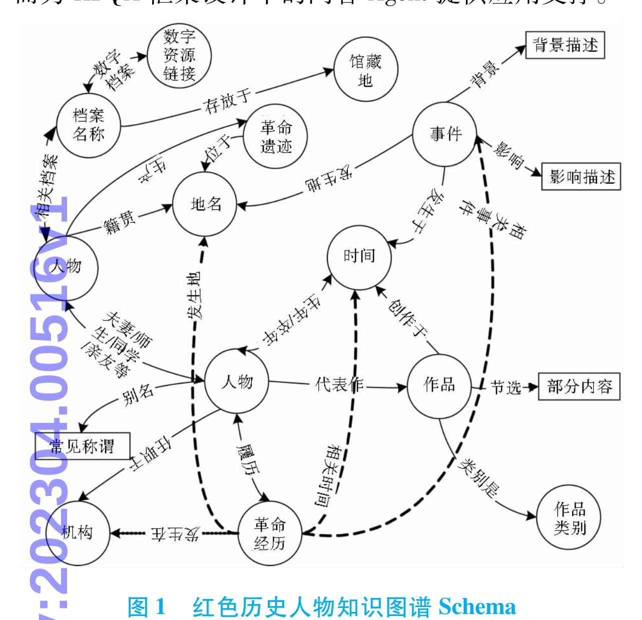


图 1　红色历史人物知识图谱 Schema

2.1.2　问答服务架构:KBQA 架构设计及其运行机制

KBQA 架构是知识问答服务架构的典型形式,具有可解释性强、部署简单、落地快速等优势。本研究设计的红色历史人物知识问答服务框架也采用此架构,主要包括了四个要素:问题、问答 agent、知识库和答案,见图 2。问题模块的任务是通过 flask 框架析取用户在聊天框页面输入的包含特定提问意图的自然语句,并将其传递给问答 agent;问答 agent 作为问答服务的核心处理框架,囊括了从识别问题到给出答案的一系列处理过程,其生成的答案通过 flask 框架反馈到答案模块,以回答特定意图的问题;答案经过问答核心组件的处理流程后,从网页聊天框回复给用户;知识库是以 NEO4J 为存储工具的知识图谱,为整体架构提供数据基础。

从运行机制上讲,问答交互方式主要以聊天框一问一答的形式进行,但其核心流程主要在"问答 agent",主要包括自然语言理解、知识图谱查询和自然语言合成三个部分。

(1)自然语言理解。自然语言理解的目标是将文

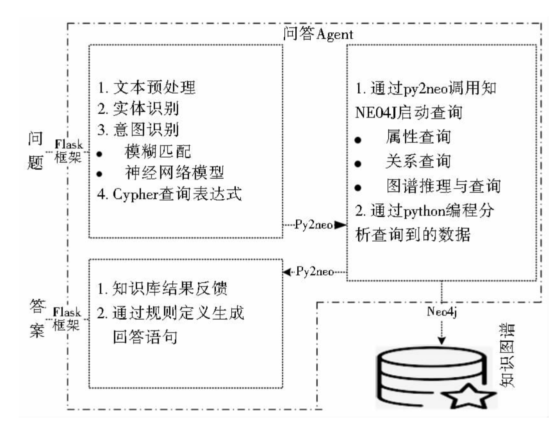


图 2　智能问答服务架构

本信息转换为可被机器处理的语义表示,在本研究中主要指问答系统需要识别用户语句中所包含的问答意图并转换为相应的查询语句。

首先需要对用户输入的自然语言文本进行预处理,主要过程包括自动分词、词性标注和去停用词。计算机无法直接对文本格式的数据进行处理,因而需要将文本转换为向量。实体识别是指能够从用户语句中自动提取含有特定意义的词语,比如人名、地名等。意图识别是指能够根据用户提出的直接或者间接信息快速判断用户的真实意图,并将对话意图与具体问题类型对应起来。意图识别本身是一个分类问题,常用的方法有基于模糊匹配和基于深度学习神经网络模型等。

(2)知识图谱查询。通过对用户输入的核心信息进行判断,分析出用户想要询问的问题后,就可以根据意图识别算法模型自动生成 Cypher 格式的数据库查询表达式,在图数据库中进行属性查询、关系查询或者知识推理。部分知识查询实例如表 1 所示:

表 1　部分问答实例知识图谱查询表

查询类别	查询实例	Cypher 查询表达式
属性查询	《作品》的作者是谁	code1 = 图谱名称. run("MATCH (n:Literature{name:'《作品名称》'}) return n"). data()
关系查询	"某人"的籍贯是哪里	code2 =图谱名称. run("MATCH (d:人物) -[:'籍贯']- >(n) WHERE d. Name = '人物名称' return n"). data()
知识推理	分析"某人"的革命履历	for i in per: print('正在分析' + data["person"] + "的" + i) sqls = "MATCH (d:人物名称) -[:{}] - >(n) WHERE d. Name = '{}' return n". format(i,data["person"]) code = 图谱名称. run(sqls). data()

需要说明的是,本研究知识推理的实现,融合了关系路径(Path Ranking Algorithm, Path Ranking)推理与

知识嵌入式表示(Translating Embedding, TransE)两种算法,通过 TransE 算法将知识图谱中的实体和关系映射到低维稠密的空间中,将 Path Ranking 推理转化为实体与关系所关联的向量或矩阵之间的运算,这种运算的操作花销比传统关系路径推理要小很多,故而能显著提升其推理效率;同时,通过对实体信息进行语义挖掘,能够搭建新的图关系路径,有助于发现隐含知识。

(3)自然语言生成。自然语言生成的任务是在正确理解用户意图的基础上,结合在知识图谱中查询到的结果,重新组织语言,以流畅、通顺、易懂的语句回答用户,自然语言生成方法通常包括检索式和生成式两种。前者依据意图类别在知识库中检索相应的答案,再利用不同的规则模板来完成语句加工和生成,优势在于生成的答案较为准确,缺陷是当识别不到问题意图时无法返回答案;后者通过大量已标注数据,训练问题－答案的神经网络模型,例如 seq2seq、attention + BILSTM、BERT 等,将问题输入到已训练好模型,可以端到端的直接返回答案语句,优势在于有问必答且答案语句多样化,但答案准确性依赖于机器学习的效果。考虑到生成式方法的答案严谨性较差,具有不确定性,不适合于红色历史人物知识问答,本研究采用检索式实现答案生成。

2.2 模型与流程

根据红色历史人物数字资源的特点,结合知识图谱构建的一般方法,本研究遵从简洁、科学、有效的原则,建立了红色历史人物数字资源知识问答模型,如图 3 所示:

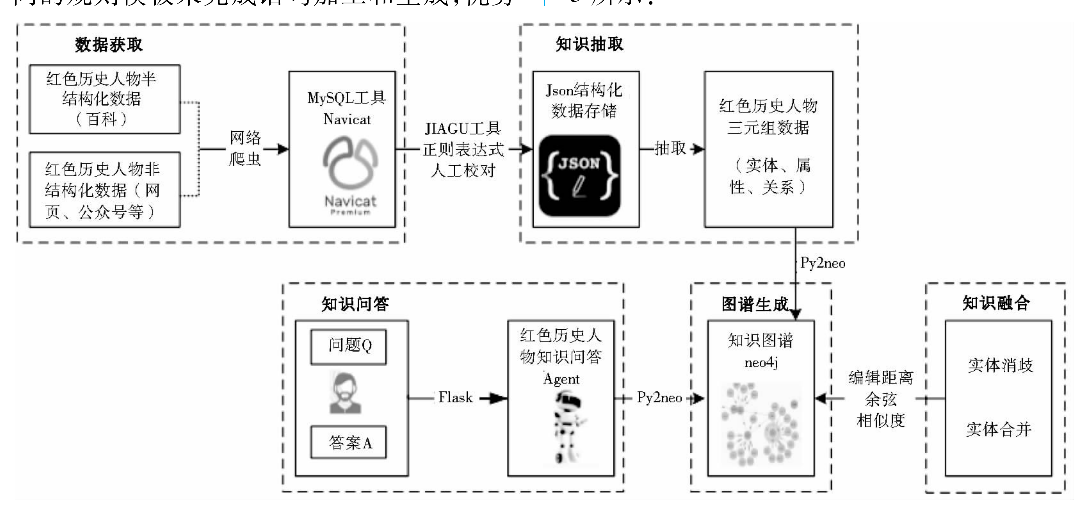


图 3 红色历史人物数字资源知识问答模型

2.2.1 数据获取

红色历史人物数字资源来源比较广泛,通常包括结构化数据、半结构化数据与非结构化数据。现存红色历史人物数字资源中,完整可用的结构化数据较为缺少,通常为半结构化数据与非结构化数据。通常,非结构化、半结构化数据可通过数据抽取技术与工具,转为格式化结构。本研究主要采用 python 爬虫,利用正则表达式对微信公众号、百度人物百科中的半结构化数据进行自动抽取,并将抽取数据转换成结构化的 JSON 格式为下一步知识抽取任务提供基础。

2.2.2 知识抽取

知识抽取是指对来源不同、结构不同的数据进行处理,抽取项目中所需要的信息形成知识,并按一定的格式将其进行存储。本文根据前文所构建的红色历史人物知识图谱 Schema 模型的信息需求对红色历史人物数字资源进行知识抽取,主要包括实体识别、属性抽取和关系抽取。

实体抽取就是对具有特定意义的实体进行抽取,主要包括人物姓名、重大事件、代表著作和地名等信息;属性抽取通常是对人物、作品或事件等实体的属性描述进行提取;关系抽取通常是以三元组的形式进行,负责提取实体间的关系并形成知识网络。本文主要采用基于神经网络与规则相结合的模型进行关系抽取,模型融合的优势在于能最大程度地解析来自不同数据源的半结构化数据与非结构化数据,为构建知识图谱奠定数据基础。

2.2.3 知识融合

知识抽取完毕后,还需采用知识融合方法,对抽取结果进行整合,通过合并红色历史人物数字资源中存在的化名、别名、地名称呼等,来完成实体消歧,为后续知识图谱的推理提供底层支撑。

文本相似度计算是知识融合常见的方法之一,其

原理是将实体中文本相似度高于一定阈值的实体进行合并。本研究对比了 Jacard 相似度、编辑距离、欧氏距离、simhash 算法、余弦相似度和 TF-IDF 等短文本相似度常用算法,以简洁有效为原则,结合红色历史人物资源相关实体特征,最终选用实体字符串相似度计算方案——加权编辑距离算法和 TF-IDF 算法。

2.2.4　图谱生成

知识融合之后的实体、属性及关系可用 RDF 三元组进行表示。RDF 三元组资源描述框架可以有效揭示数据与数据之间的联系,其序列化的方式主要包括 RDF/XML、N-Triples、Turtle、RDFa,JSON-LD 等。本文采取的方案是通过 JSON-LD 以键值对的方式形象地存储三元组数据,再通过 python 程序语言中的 py2neo 第三方库将三元组知识存储到图数据库 NEO4J 中,该方案的优势是响应快速、兼容性强、易于落地。

2.2.5　知识问答

知识图谱构建完成后,可以在此基础上实现红色历史人物知识智能问答。本研究主要通过自然语言理解、知识查询、自然语言生成三个主要步骤实现问答过程。

在问答设计过程中,借鉴了多轮问答中意图识别与槽填充的思想,结合机器学习算法完成自然语言识别的功能。在将意图识别结果自动转换成 Cypher 表达式后,通过调用 python 的 py2neo 库在 NEO4J 中完成知识查询,最后对结果进行解析并生成答案。

3　实证研究:老上大红色历史人物知识问答服务

3.1　对象选择

本文选取老上大历史人物数字资源作为案例进行实证研究。一方面,从历史角度,老上大历史人物的活跃时期多处于二十世纪二十年代,属于早期的红色历史人物,其历史贡献卓著,与红色历史活动的发源地上海密切相关;另一方面,研究团队从 2014 年起,致力于对老上大历史人物资料展开收集和整理,并建立公众号“上大故事”,以规范的数据板块对 52 位老上大历史人物知识展开推介,数据形式为非结构数据,而作为半结构化数据的百度百科的知识则可作为本研究的补充数据源。综上,从历史视角和数据视角考虑,选择老上大历史人物作为案例进行红色历史人物知识问答服务框架实证具有一定的代表性和可操作性。

3.2　关键环节

3.2.1　老上大红色历史人物数据获取

通过对“上大故事”公众号发布的 52 位人物专题进行网络爬虫,获取人物相关文章共 146 篇,部分爬虫程序如图 4 所示。为了使内容更加完整全面,本研究从百度百科人物简介获取了一些人物基本信息和革命履历描述内容作为补充。

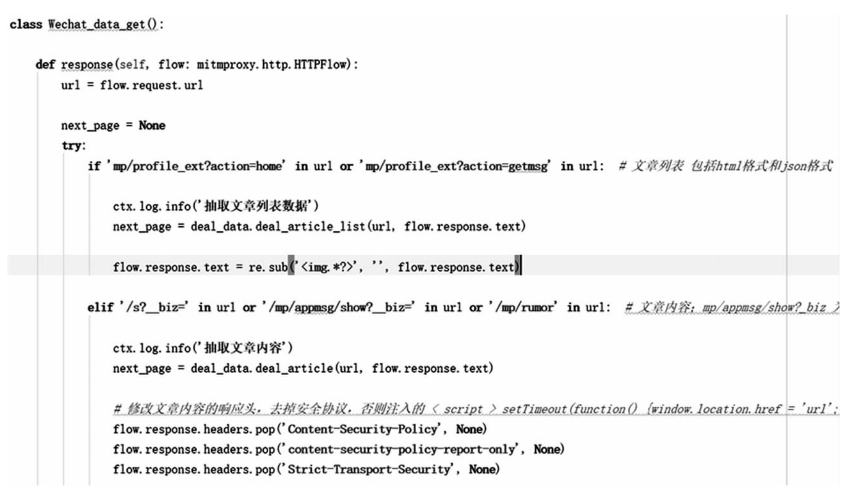


图 4　公众号爬虫部分代码截图

3.2.2　老上大红色历史人物知识抽取

鉴于选取案例缺乏有关老上大红色历史人物数据的专有名词标注数据集,本文采取规则与深度学习神经网络模型相结合的方式来进行知识抽取。实体抽取主要通过爬虫分析、结巴分词专有名词的自动提取来实现;属性和关系主要以三元组关系进行抽取,具体方式包括在网络爬虫时根据半结构化数据的 H5 标签进行自动提取,利用 Jiagu 深度学习神经网络开源模型对段落篇章等非结构化数据进行自动化抽取。Jiagu 以 BILSTM 模型等为基础,使用大规模中文语料训练而

成，由于已存在预训练模型，其使用时无需额外标注，可通过调用 python 第三方库相关函数实现知识抽取功能。

例如，输入“瞿秋白 1899 年 1 月 29 日出生于江苏常州，本名双，后改瞿爽、瞿霜，字秋白，生于江苏常州，中国共产党早期主要领导人之一”时，模型将自动就句子中所含的三元组关系进行抽取罗列，结果如下图 5 所示：

```
E:\anaconda3\envs\py36\lib\site-packages\tensorflow\python\framework\dtypes.py:525: FutureWarning: Passing (type, 1) or '1type' as a synonym
  np_resource = np.dtype([("resource", np.ubyte, 1)])
E:\anaconda3\envs\py36\lib\site-packages\tensorboard\compat\tensorflow_stub\dtypes.py:541: FutureWarning: Passing (type, 1) or '1type' as a
  _np_qint8 = np.dtype([("qint8", np.int8, 1)])
E:\anaconda3\envs\py36\lib\site-packages\tensorboard\compat\tensorflow_stub\dtypes.py:542: FutureWarning: Passing (type, 1) or '1type' as a
  _np_quint8 = np.dtype([("quint8", np.uint8, 1)])
E:\anaconda3\envs\py36\lib\site-packages\tensorboard\compat\tensorflow_stub\dtypes.py:543: FutureWarning: Passing (type, 1) or '1type' as a
  _np_qint16 = np.dtype([("qint16", np.int16, 1)])
E:\anaconda3\envs\py36\lib\site-packages\tensorboard\compat\tensorflow_stub\dtypes.py:544: FutureWarning: Passing (type, 1) or '1type' as a
  _np_quint16 = np.dtype([("quint16", np.uint16, 1)])
E:\anaconda3\envs\py36\lib\site-packages\tensorboard\compat\tensorflow_stub\dtypes.py:545: FutureWarning: Passing (type, 1) or '1type' as a
  _np_qint32 = np.dtype([("qint32", np.int32, 1)])
E:\anaconda3\envs\py36\lib\site-packages\tensorboard\compat\tensorflow_stub\dtypes.py:550: FutureWarning: Passing (type, 1) or '1type' as a
  np_resource = np.dtype([("resource", np.ubyte, 1)])
WARNING:tensorflow:From E:\anaconda3\envs\py36\lib\site-packages\jiagu\bilstm_crf.py:28: The name tf.GraphDef is deprecated. Please use tf.c

WARNING:tensorflow:From E:\anaconda3\envs\py36\lib\site-packages\jiagu\bilstm_crf.py:40: The name tf.Session is deprecated. Please use tf.cc

[['瞿秋白', '出生日期', '1899年1月29日'], ['瞿秋白', '出生地', '江苏常州'], ['瞿霜', '出生地', '江苏常州'], ['瞿霜', '国籍', '中国']]

Process finished with exit code 0
|
```

图 5　知识抽取程序运行示意

本研究使用以上方法半自动化抽取实体数量 531 个、属性 67 个、三元组数量 421 个，后续通过人工校对和补充增加实体 52 个、属性 2 个、三元组 39 个。具体知识抽取情况如表 2 所示：

表 2　知识抽取结果统计表

知识来源	实体数量	属性数量	三元组数量
Python 爬虫	289	35	191
Jieba 词库	62	0	0
Jiagu 模型	180	32	230
人工校对新增	52	2	39
共计	583	69	460

3.2.3　老上大红色历史人物实体融合

本文采用加权编辑距离算法和 TF-IDF 两个算法来计算实体字符串的相似度。

编辑距离（Levenshtein Distance）是 NLP 中一种计算两个字符串间差异程度的字符串度量指标，其主要特点在于使用了动态规划的思想来比较文本结构，在短文本相似度计算上速度快且准确率较高。两个红色历史人物数字资源实体字符串 a，b 的 Levenshtein Distance 可表示为 $lev_{a,b}(|a|,|b|)$，其中 |a|，|b| 分别对应 a，b 的长度。通过矩阵计算和循环迭代，可得出两个实体字符串的编辑距离并记为 L_Distance，这里的 i，j 在计算时可视为是 a，b 的长度 |a|，|b|，计算方式如公式（1）所示：

$$lev_{a,b}(i,j)=\begin{cases}\max(i,j) & \text{if } \min(i,j)=0\\ \min\begin{cases}lev_{a,b}(i-1,j)+1\\ lev_{a,b}(i,j-1)+1\\ lev_{a,b}(i-1,j-1)+1_{(a_i\neq b_j)}\end{cases} & \text{otherwise}\end{cases}$$

公式（1）

TF-IDF 是一种统计方法，通过 TF-IDF 算法提取的词向量能较好地反映实体字符串之间的差异性。在提取得到两个实体字符串的词向量之后，通过向量余弦公式可以计算出他们的相似度值 $sim(a,b)$，记为 T_Distance，计算方式如公式（2）所示：

$$sim(a,b)=\cos(\theta)=\frac{a*b)}{\|a\|\ \|b\|}=\frac{\sum_{k=1}^{n}a_k\times b_k}{\sqrt{\sum_{k=1}^{n}a_i^2}\times\sqrt{\sum_{k=1}^{n})\ b_i^2}}$$

公式（2）

经过实验调优，最终两个实体字符串的相似度如公式（3）所示：

$$similarity=0.6*L_Distance+0.4*T_Distance$$

公式（3）

当 *similarity* 相似度值大于 0.85 时就认为 a，b 两个字符串属于同一实体，即可通过 Cypher 语言的 merge 函数进行相同实体合并，否则认为 a，b 属于不同实体。通过以上知识融合步骤，本研究共融合实体 236 个，完成了对多源实体信息的合并融合，便于系统囊括多种渠道数据丰富三元组知识，有效保障了后续推理和问答的准确率。

3.2.4　老上大红色历史人物图谱生成

将三元组数据整理完成之后，利用 python 的第三方库 py2neo，通过 Cypher 语句将三元组数据自动导入到图数据库 NEO4J 中，即可依托 NEO4J 可视化功能展示老上大红色历史人物可视化知识图谱。知识图谱中共有实体 347 个，属性数量 69 个，关系数量 460 个，部分可视化界面如图 6 所示：

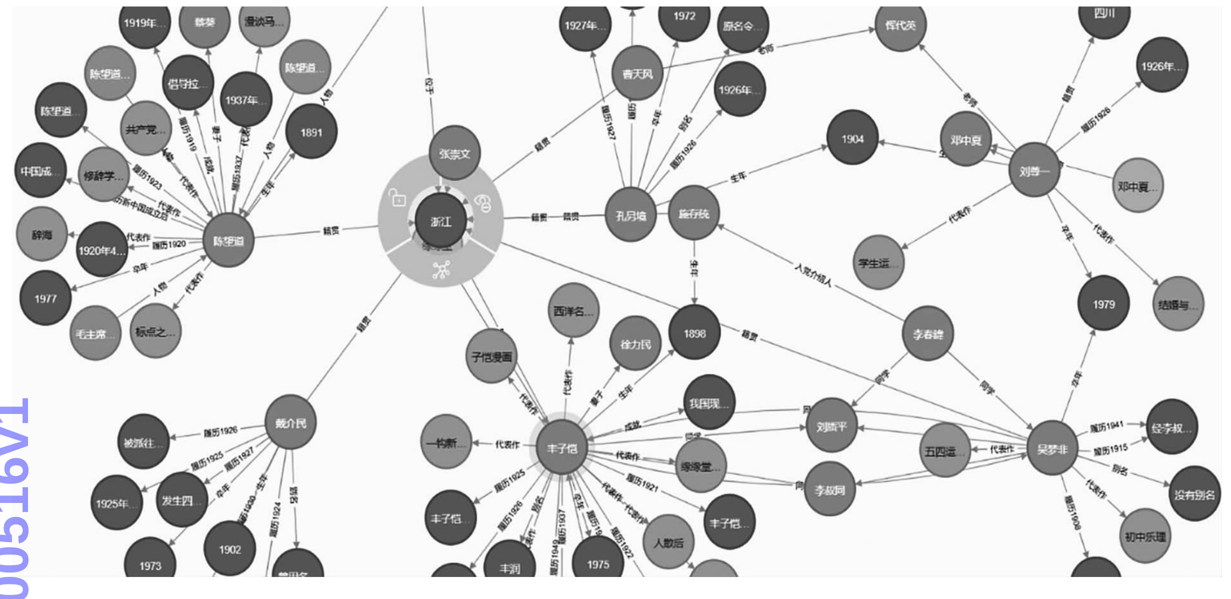


图 6　老上大历史人物 NEO4J 部分数据截图

从图 6 中可以直观看出部分老上大历史人物的三元组信息。例如，施存统、陈望道、丰子恺和张崇文等人的籍贯都是浙江省；李叔同是丰子恺和吴梦非的老师；陈望道的代表作之一是共产党宣言的翻译著作，创建于 1920 年，从代表作节选的实体属性中还能看到该著作的内容节选；缘缘堂是与丰子恺相关的遗迹，位于浙江。

3.2.5　老上大红色历史人物知识问答

（1）文本预处理。本案例采用开源工具结巴分词来对问句进行自动分词、词性标注。通过结巴分词内置的自定义词典功能，将老上大历史人物知识图谱中的实体通过 python 编程语言创建到结巴自定义词典，实现了对用户问句中所含实体的自动抽取，为意图识别奠定了基础。

（2）意图识别。本案例采用槽填充和朴素贝叶斯算法构建了意图识别的主要模型，并辅助模糊匹配算法对意图较为不明确的语句进行二次识别。

“槽填充 + 机器学习分类算法”是一种可精确识别用户问题类别的意图识别方法。槽是对话过程中将初步的用户意图转化为明确的用户指令所需补全的信息[19]。一个槽与一次问答处理中所需要获取的某个信息相对应，回答完整问句通常需要基于实体、属性、关系构建链式槽位并做相应填充。以问句是“瞿秋白 1937 年在什么单位任职”为例，可构建如下链式槽，槽位识别后将自动填充为{‘entity_num’:[1],‘entity’:[‘瞿秋白’],‘quality_deep’:[‘True’],‘quality’:[‘position’],‘intent’:[]}。

槽填充完成后，需要将槽位填充的属性与具体问题对应起来，此功能需要朴素贝叶斯算法来辅助实现。针对不同槽位依次填充的结果，采用朴素贝叶斯算法学习意图识别模型的输入和输出的联合概率分布，并求出后验概率最大的输出，即为对话最可能的意图。如果用户语句所含信息不足以填充足够槽位进而完成意图分类，则需要用到模糊匹配作为辅助模型。本案例以编辑距离和余弦相似度算法相加权作为短文本相似度计算的核心步骤，对计算结果按降序排列取最高概率值，即为最可能的问答意图。

3.3　结果展示

3.3.1　系统结构

本案例搭建的智能问答系统原型框架如图 7 所示，主要包括数据获取、数据处理、图谱生成、意图识别、sql 语句生成、答案查询、前后端 flask 交互等环节。

系统原型的前端采用 H5 搭建了一个网页聊天框式的服务；后端采用 python 编程语言来实现整个智能化识别、查询、生成的底层功能；而前后端交互则采用了目前十分流行的 Flask web 框架。Flask 框架的主要特征是核心构成比较简单，但具有很强的扩展性和兼容性，因此可以快速实现一个网页智能聊天的服务。

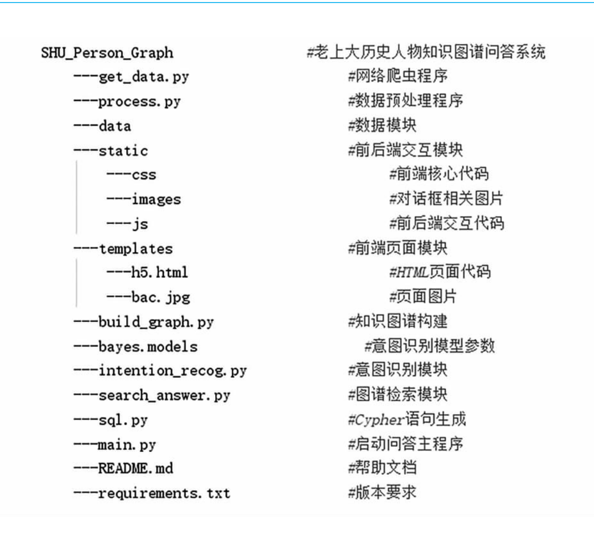


图7 老上大历史人物知识图谱问答系统架构

3.3.2 问答示例

本案例的问答示例主要涉及红色历史人物的基本信息、社会关系、革命经历、作品著述、数字档案、智能推理等典型问题，图8展示了问答系统的交互过程。其中，基本信息问答展示了有关瞿秋白、吴梦非和马宁等人籍贯、出生年份、个人成就等方面的问答；社会关系问答展示了“许心影的老师”和“王环心的入党介绍人”等定向社会关系问答，还有“吴梦非社会关系都有哪些”的遍历式社会关系问答；革命经历问答展示了柯柏年、施存统、丰子恺三位老上大历史人物在不同年份革命履历的问答；作品著述问答展示了对于陈望道及其著作信息的问答，如“陈望道有哪些代表作”“共产党宣言是什么类型的作品”“给我推荐一下共产党宣言的精彩片段”等；数字档案问答展示了部分人物数字档案名称、多媒体链接、存放地址和位置等的问答；智能推理问答中展示了“老上大历史人物中与五卅运动相关的都有哪些人”“参加了开国大典的老上大人物都有哪些”“籍贯是浙江的都有谁”等问题的回答。



图8 问答系统知识问答案例演示

从以上问答实例可以看出，问答系统原型能较准确地识别用户自然语言的问题意图，通过对图数据库中不同实体的内容进行分析、相关节点进行遍历与追溯，还能针对较复杂的问题利用知识图谱进行推理求解，从而对典型红色历史人物的问题较好地完成了解答。

3.3.3 问答测试

本研究对上文所实现的问答服务做了问答准确

率、应答速度两方面的系统功能测试。研究采集了老上大历史人物知识问答爱好者的热门问题 200 余条，再通过人工校对、分类、补充，制作成主题囊括基本信息、社会关系、革命经历、作品著述、数字档案和智能推理六类，共 150 个问答测试对（含问题与标准答案）作为测试数据集。经测试，本系统平均准确率达到 90.6%。实验证明，本研究开发的自动问答系统可以较为准确的回答大多数问题，具体测试结果如表 3 所示：

表 3　问答系统测试结果统计表

问题类别	测试问题数量/个	回答准确数量/个	回答准确率/%
基本信息	25	24	96
社会关系	25	22	88
革命经历	25	24	92
作品著述	25	24	96
数字档案	25	23	92
智能推理	25	19	76
共计	150	136	90.6

同时，通过使用 ApacheJMeter 工具对问答系统的速度性能进行了 200 个并发量的测试，结果如图 9 所示。对话服务平均延迟为 0.16 秒，故可满足百人以上规模同时向问答系统发送请求并迅速做出回答的场景，但是对于区域更广、人数更多时的应用落地，延迟达 0.5 秒以上，无法快速做出响应和回答，性能仍然存在不足。

3.4　结果分析与讨论

本研究利用知识图谱对老上大历史人物相关数字资源进行了描述、组织与关联，并在此基础上实现了基于知识图谱的智能问答系统原型。该系统的优越性体现在以下几点：

（1）实现可视化展示，革新资源组织方式。提供了从半结构化与非结构化数据中整合红色历史人物资源、构建知识图谱的技术方案，并依托图数据库实现了红色历史人物关键信息资源的存储和可视化。

（2）依托图谱化关联，提升回答准确率。以知识图谱为基础的对话系统擅长于解决垂直领域的问答，老上大历史人物知识问答系统即属于此类，其实现了对关联数据的智能问答，准确性高于闲聊问答机器人。

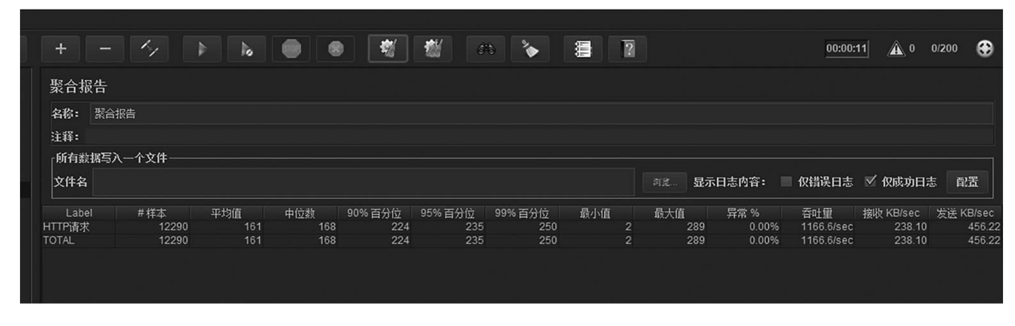


图 9　系统性能测试截图

（3）实现智慧化交互，推广应用前景良好。老上大历史人物知识问答系统可以封装成微信小程序面向更为广阔的受众群体，可大大提高红色历史人物数字资源的普及程度，激发广大读者的学习兴趣，为红色知识的推广传播提供一种新的途径。

同时，实证案例也发现了研究中存在的诸多不足：受老上大历史人物范围所限，人物样本数量较少；缺少足够的已标注的红色历史人物训练集，自动化抽取需要辅以人工校对；并发量超过 200 时速度性能不足；问答展示的应用形式不够多元化等等，这些都是本研究后续要推进解决的问题。

4　结语

数字人文背景下大数据、人工智能等技术的不断成熟和深入应用，改变了传统的知识组织和服务方式，有效地利用这些技术，将为红色数字资源的知识组织和服务变革培植新的动力。本文实现的老上大历史人物的知识问答系统，提供了从多源数据中提取红色历史人物关键数据并构建知识图谱的通用方案，并在知识图谱构建的基础上，探析了红色历史人物知识问答系统的典型应用，为知识图谱技术赋能知识组织及知识服务模式提供了新思路，更为红色数字资源深度开发利用提供了新路径。

本研究的主要创新聚焦于红色历史人物数字资源 KBQA 架构设计及其运行机制：①结合红色历史人物知识的特点，借鉴多轮问答中意图识别的槽填充方案，提出"槽填充 + 机器学习分类算法"的意图识别方法，提高了意图识别的精度；②采用 TransE 和 Path Ranking 相结合的算法，实现了基于知识图谱的智能推理。下一步，本研究将一方面扩展红色历史人物样本集，探

索准确率更高、速度更快的模型来规模化建构红色历史人物知识图谱，奠定坚实的数据基础设施，另一方面，本研究将采用微信小程序、APP 等多元化形式拓宽红色数字资源知识服务应用渠道。

参考文献:

[1] 中共中央办公厅,国务院办公厅. 关于实施革命文物保护利用工程(2018－2022 年)的意见[N]. 人民日报,2018－07－30(001).

[2] Harvard University, Academia Sinica, Peking University. China biographical database [EB/OL]. [2021－03－03]. https://projects.iq.harvard.edu/cbdb.

[3] 张文勇. 湖南近代人物数据库建设研究与实现[D]. 长沙:中南大学,2013.

[4] 靳志军,何寿峰,郝春柳,等. 河北红色历史文化资源挖掘研究－以李大钊专题数据库建设为例[J]. 文化月刊,2016(7):114－115.

[5] 刘宁静,刘音,王莫言,等. 数字人文视角下学术名人知识模型构建研究——以李政道数字资源中心为例[J]. 图书情报工作,2019,63(23):113－121.

[6] 陈涛,刘炜,单蓉蓉,等. 知识图谱在数字人文中的应用研究[J]. 中国图书馆学报,2019,45(6):34－49.

[7] 潘俊. 面向数字人文的人物分布式语义表示研究——基于 CBDB 数据库和古籍文献[J]. 图书馆杂志,2020,39(8):94－102.

[8] 杨海慈,王军. 宋代学术师承知识图谱的构建与可视化[J]. 数据分析与知识发现,2019,3(6):109－116.

[9] 周亦,周明全,王学松,等. 大数据环境下历史人物知识图谱构建与实现[J]. 系统仿真学报,2016,28(10):2560－2566.

[10] 单良,刘欣. 基于中国历史人物知识的智能问答系统构建[J]. 情报探索,2019(6):101－105.

[11] COCCO R, ATZORI M, ZANIOLO C, etal. Machine learning of SPARQL templates for question answering over LinkedSpending [J]. CEUR workshop proceedings,2019,2400:156－161.

[12] 李贺,刘嘉宇,李世钰,等. 基于疾病知识图谱的自动问答系统优化研究[J]. 数据分析与知识发现,2021,5(5):115－126.

[13] 陈璟浩,曾桢,李纲. 基于知识图谱的"一带一路"投资问答系统构建[J]. 图书情报工作,2020,64(12):95－105.

[14] 高劲松,方晓印,刘思洋,等. 基于关联数据的馆藏文物资源知识关联与智能问答研究[J]. 情报科学,2021,39(5):12－20.

[15] WU W Q, ZHU Z F, LU Q, etal. Introducing external knowledge to answer questions with implicit temporal constraints over knowledge base[J]. Future Internet,2020,12(3):45.

[16] 谢项. 基于古诗知识图谱的智能问答研究[D]. 武汉:华中师范大学,2020.

[17] YU J, ZHU Z H, WANG Y J, etal. Cross-modal knowledge reasoning for knowledge-based visual question answering [EB/OL]. [2021－08－02]. https://www.sciencedirect.com/science/article/pii/S0031320320303666? via%3Dihub.

[18] HUANG T S, LI X W, ZHAI S P, etal. Knowledge graph reasoning based on tensor decomposition and MHRP-Learning [EB/OL]. [2021－08－02]. https://downloads.hindawi.com/journals/am/2021/8880553.pdf.

[19] TANG H, DONG H J, ZHOU Q J. End-to-end masked graph-based CRF for joint slot filling and intent detection[J]. Neurocomputing, 2020,413(6):348－35.

作者贡献说明:

张云中:确定选题,提出研究思路,设计研究方案,论文修改;

郭冬:程序设计,实验实施,论文初稿撰写及修改;

王亚鸽:数据处理和论文修改;

孙平:数据处理和论文修改。

Framework of Knowledge Q & A Service for Red Historical Figures Based on Knowledge Graph

Zhang Yunzhong Guo Dong Wang Yage Sun Ping

Department of Library, Information and Archives, Shanghai University, Shanghai 200444

Abstract: [**Purpose/significance**] Knowledge graph has become a new form of public digital cultural resources organization. Using knowledge graph technology to enable the Knowledge Q & A service of red historical figures and improve user interaction experience is of great significance to the development and utilization of red historical resources. [**Method/process**] On the basis of combing the related research of digital resource organization and Knowledge Q & A system of historical figures, the paper constructed the knowledge graph schema and KBQA architecture of red historical figures, and then built the model of Q & A from five aspects of data acquisition, knowledge extraction, knowledge fusion, graph generation and Knowledge Q & A. This paper took the red historical figures digitalresources of Shanghai University(1922－1927) as an example for empirical research. [**Result/conclusion**] The Knowledge Q & A service architecture designed in this paper has advantages in semi-automatic graph construction, knowledge reasoning and intelligent interaction of digital resources of red historical figures, and improves the user knowledge service experience.

Keywords: red historical figures knowledge graph question answering system knowledge service